Plan of Action to Prevent Human Extinction Risks

Abstract:

During the last few years a significant number of global risks have been discovered that threaten human existence. These include, to name but a few, the risk of harmful AI, the risk of genetically modified viruses and bacteria, the risk of uncontrollable nanorobots-replicators, the risk of a nuclear war and irreversible global warming. Additionally, dozens of other less probable risks have been identified. Also a number of ideas have been conveyed regarding the prevention of these risks, and various authors have campaigned for different ideas.

This roadmap compiles and arranges a full list of methods to prevent global risks. The roadmap describes plans of action A, B, and C, each of which will go into effect if the preceding one fails.

Plan A is prevent global risks, it combines 5 parallel approaches: international control, decentralized monitoring, friendly AI, rising resilience and space colonization.

Plan B is to survive the catastrophe.

Plan C is to leave traces.

Plan D is improbable ideas.

Bad plans are plans that raise the risks.

The document exists in two forms: as a visual map (pdf http://immortality-roadmap.com/globriskeng.pdf) and as a text (long read below – 50 pages).

Introduction

The problem

Many authors noted that in the 21th century may witness a global catastrophe caused by new technologies (Joy, Rees, Bostrom, Yudkowsky, etc).

Many of them suggested different ways of x-risks prevention (Joy, Posner, Bostrom, Musk, Yudkowsky).

But these ideas are disseminated in literature and unstructured, so we need to collect all of them, put them in a most logical order and evaluate their feasibility.

As a result, we will get a most comprehensive and useful plan of x-risks prevention that may be used by individuals and policymakers.

In order to achieve this goal I created a map of x-risks prevention methods.

The map contains all known ways to prevent global risks, most of which you may have probably heard of separately.

The map describes the action plans A, B, C, D, each of which will come into force in the event of the failure of a previous one. The plans are plotted vertically from top to bottom. The horizontal axis represent timeline and some approximate dates when certain events on the map may occur.

The size of this explanatory text is limited by the size of the article, so many points on the map I left as self-evident or linked them to explanations by other authors. A full description of every point would take up a whole book.

The context

The context of the map is an exponential model for the future. The map is based on the model of the world in which the main driving force of history is the exponential development of technology, and in which a strong artificial intelligence will have been created around 2050. This model is similar to the Kurzweil model, although the latter suffers from hyper-optimistic bias and does not take account of global risks.

This model is relatively cautious as compared to other exponential models, for example, there are models where technology development takes place according to a hyperbolic law and there is a singularity around 2030 (Scoones, Vinge, Panov, partly Forester).

At the same time we must understand that this model is not a description of reality, but a map of a territory, that is, in fact, we do not know what will happen, and very serious deviations are possible because of black swan events or through slower technological growth.

I should note that there are two other main models – the standard model, in which future will be almost as today with a slow

linear growth (this model is used by default in economic and political forecasting, and it is quite good over intervals of 5-10 years) and the model of Rome Club, according to which in the middle of the 21st century there will be a sharp decline in production, economy and population. Finally, there is the model of Taleb (and Stanislaw Lem), in which the future is determined by unpredictable events.

In fact, we don't have a good plan

The situation is that in fact we do not have a good plan, because each plan has its own risks, and besides, we do not know how these plans could be implemented.

That is, although there is a large map of risk prevention plans, the situation of prevention does not look good. It is easy to criticize each of the proposed plans as unrealizable and dangerous, and I will show their risks. Such criticism is necessary for improving the existing plans.

But some plan is better than no plan at all.

Firstly, we can build on it to create an even better plan.

Secondly, the mere implementation of this plan will help delay a global catastrophe or reduce its likelihood. Without it, the probability of a global catastrophe is estimated by different scientists at 50 per cent before the end of the 21st century.

I hope the implementation of a most effective x-risks prevention plan will lower it by order of magnitude.

Overview of the map

Plan A "Prevent the catastrophe" is composed of four sub-options: A1, A2, A3 and A4. These sub-options may be implemented in parallel, at least up to a point.

The idea of plan A is to completely avoid the global catastrophe and to achieve such a state of civilization, that its probability is negligible. The sub-options are following:

- Plan A1 is the creation of a global monitoring system. It includes two options: A1.1 –international centralized control and A1.2 – decentralized risk monitoring. The first option is based on suppression, the second is co-operative. The second option emerged during crowdsourcing ideas for the map in summer 2015.
- Plan A2 is the creation of Friendly AI,
- Plan A3 is increasing resilience and indestructibility
- Plan A4 is space colonization,

Among them the strongest are the first two plans and in practice they will merge: that is, the government will be computerized, and AI will take over the functions of the world government.

- Plan B is about building shelters and bunkers to survive the catastrophe,
- Plan C is to leave traces of information for future civilizations.

- Plan D is hypothetical plans
- "Bad plans" are dangerous plans that are not worth implementing.

The procedure for implementing the plans

In order to build a multi-level protection against global risks, we should implement almost all of the good plans. At early stages, most plans are not mutually exclusive.

The main problem that can make them begin to exclude each other, arises in connection with the question of who will control the Earth globally: a super UN, AI, a union of strong nations, a genius hacker, one country, or a decentralized civil risk monitoring system. This question is so serious that in itself is a major global risk, as there are many entities eager to take power over the world.

The ability to implement all the listed plans depends on the availability of sufficient resources. Actually, the proposed map is a map of all possible plans, from which one may choose for implementation one most suitable sub-group.

If resources are insufficient, it may make sense to focus on one plan only. But who will choose?

So here arises a question of actors: who exactly would implement these plans? Currently, in the world there are many independent actors, and some of them have their own plans to prevent a global catastrophe. For example, Elon Musk proposes to create a safe AI and build a colony on Mars, and such plans could de realized by one person.

As a result, different actors will cover the whole range of possible plans, acting independently, each with his own vision of how to save the world.

Although any of the plans is suitable to prevent all possible accidents, one particular plan is most efficient for a certain type of disasters. For example, Plan A1 (international control system) is best suited to control the spread of nuclear, chemical, biological weapons and anti-asteroid protection whereas Plan A2 is the best to prevent the creation of an unfriendly AI.

Space exploration is better suited to protect against asteroids but does very little to protect against an unfriendly AI that can be distributed via communication lines, or against interplanetary nuclear missiles.

The probability of success of the plans

Maps are also arranged in order of the likelihood of the success of their implementation. In all cases, however, it is not very large. I will give my evaluation of the probability of success of the plans, highest to lowest:

Most likely is the success of the international control system A1.1, because it requires no fundamental technological or social solutions that would not have been known in the past. 10 percent. (This is my estimation of the probability that the realization of this plan will prevent a global catastrophe on the condition that no other plan has been implemented, and that the catastrophe is inevitable if no prevention plans exist at all. The notion of probability for x-risks is complicated and will be discussed in a separate paper and map "Probability of x-risks".) The main factors lowering its probability are the well-known human inability to unite, risks of world war during attempts to unite humanity forcefully and risks of failure of any centralized system.

Decentralized control in A1.2 is based on new social forms of management that are a little bit utopian so its success probability is also not very high and I estimate it at 10 percent.

Creating artificial intelligence (A2) requires the assumption that AI is possible, and this plan carries its own risks, and also AI is not able to prevent the risk of accidents that can happen before its creation, such as a nuclear war or a genetically engineered virus – 10 percent.

A3: Increasing resilience and strengthening the infrastructure can have only a marginal effect in most scenarios as help in realization of other plans, so – 1 percent.

A4: Space colonization does not protect from radio-controlled missiles, nor from the a hostile AI, or even from the slow action of biological weapons which work like AIDS. Besides, space colonization is not possible in the near future, and it creates new risks: large space ships could be kinetic weapons or suffer catastrophic accidents during lunch, so – 1 percent.

Plan B is obviously less likely to succeed, since major shelters could be easily destroyed and are expensive to build, and small shelters are vulnerable. In addition, we do not know from what type of future disasters we are going to protect ourselves building shelters. So, 1 per cent.

Plans C and D have almost symbolic chances for success. 0,001 per cent.

Bad plans will increase the likelihood of a global catastrophe.

We could hope for positive integration of different plans. For example, Plan A1 is good at early stages before the creation of a strong AI, and Plan A2 is a strong AI itself. Plan A3 will help implement all other plans. And plans A4, B and C may have strong promotional value to raise awareness of x-risks.

In the next chapters I will explain different blocks of the map.

Steps

Timeline of the map consist not only of possible dates which could move by decades depending on the speed of progress and

other events, but also of steps, which are almost the same for every plan.

Step 1 is about understanding the nature of risks and creating a theory.

Step 2 is about preparation, which includes promoting an idea, funding and building an infrastructure needed for risks mitigation. Step 2 can't be done successfully without Step 1.

Step 3 is the implementation of preventive measures on a low technological level, that is on the current level of technologies. Such measures are more realistic (bans, video surveillance) but also limited in scope

Step 4 is the implementation of advanced measures based on future technologies which will finally close most risks, but which themselves may pose their own risks.

Step 5 is the final state where our civilization will attain indestructibility.

These steps are best suited to Plan A1.1 (international control system) but are needed for all the plans.

Plan A. Prevent the catastrophe

Plan A1. Super UN or international control system

The idea of this plan is that the more complex and "aggressive" a risk, the greater the level of control is required to prevent it. As a global risk can arise in one part of the world (a genetically modified virus) and then spread across the planet, the control should be spread throughout the world (and even beyond, to the space colonies).

In order to create an adequate system of control it is necessary to understand the nature of the risks, how to detect them and how to suppress them, before they have time to spread.

However, to achieve that we need a clear understanding of the importance of preventing such risks, and a world-wide authority that would be specifically created for their prevention and would have powers that go beyond those of any local authorities.

In addition, the control system must be adequate to new technological risks and evolve in parallel with them. It may be a risk in itself, and it also has to be controlled.

A1.1 – Step 1: Research

Information gathering

- Creating and promoting a long-term future model
- Comprehensive list of risks
- Probability assessment
- Prevention roadmap
- Determining most probable risks and risks that are easiest to prevent
- Creating an x-risks wiki and an x-risks internet-forum which would attract the best minds, but would also be open to everyone and well-moderated

Assistance

- Solving the problem of different x-risk-aware scientists ignoring each other (the "world savior's arrogance" problem)
- Integrating different lines of thinking about x-risks
- Lowering the barriers to entry
- Unconstrained funding of x-risk research for many different approaches
- Helping best thinkers in the field (Bostrom) produce high quality x-risk research
- Educating "world saviors": choosing best students, providing them with courses, money and tasks.

Additional study areas

- Studying existing system of decision-making in UN, hiring a lawyer
- Creating a general theory of safety and risk prevention
- Creating a full list of x-risk-related cognitive biases and working to prevent them (Yudkowsky)
- Translating best x-risk articles and books into common languages

The basis of the modern understanding of global risks has been laid in the works of Nick Bostrom, Leslie, Martin Rees, Bill Joy at the beginning of the 21st century. The output is a more or less complete list of risks.

I made a typology of the global risk map (http://lesswrong.com/lw/mdw/a_map_typology_of_human_extinct

ion_risks/) that shows more than 100 different options, but the main risks in the exponential model of the future are the risks of new technologies, namely Artificial Intelligence and a multipandemic caused by genetically modified viruses. These top two risks are growing exponentially along with the development of technology, and their probability grows at the same rate as Moore's Law, that is, doubling every couple of years. However, several other risks could also lead to a global catastrophe: a world war with nuclear-biological weapons involved, a nuclear weapons doomsday, irreversible global warming under the Venus scenario creation, nanorobots-replicators.

The issue of estimating the probability of certain risks is certainly not resolved. Partly because it is very difficult to estimate the probability of a unique single event that has never happened before. Even a notion of probability is not defined for such events. I am planning to make a map that will show time distribution of various risks.

A number of various ways to protect the global risk have recently been proposed. Yudkowsky and Bostrom favor creation of a friendly AI. Hawking and Elon Musk also advocated for creating space shelters. R. Posner in his book "Catastrophe: Risk and reaction" described what I call here Plan A1, that is the creation of the international regulatory mechanisms to prevent risks. Various options of underground shelters have been suggested.

Plan A1.1: Step 2: Social support

Public Movement: spreading the idea of the importance of risk prevention

Science

- Rising support inside academic community by high level signaling
- Cooperative scientific community with shared knowledge and productive rivalry
- Productive cooperation between scientists and society based on trust
- Scientific attributes:
  - a peer-reviewed journal,
  - conferences,
  - an inter-governmental panel,
  - an international institute

Popularization

- Articles, books, forums, media: showing the public that x-

risks are real and diverse, but we could and should act

Politics

- Public support, street action (anti-nuclear protests in 80s)
- Political support: lobbyism
- Establishing political parties for x-risk prevention
- Writing policy recommendations

Unfortunately, the scientific community tends to split into opposing groups. As a result, one group focuses on a single risk and particular method of its prevention while another group concentrates on other risks and methods. (E.g. , global warming and CO2 reduction as a method of its prevention.) But active influence on policymakers requires a team of scientists with a common (and correct) vision. Now one such group has emerged around Yudkowsky – Bostrom – Elon Musk, but they seem to overestimate remote risks of superintelligence and to underestimate other risks that could happen earlier.

In general, we live in a society that avoids solving the biggest problems and focuses on small issues. The same thing happens with the fight against aging, the number one killer in the world.

The second step is to convince the society and decision-makers in the reality of the threat of global risks and the need to deal with them as the humanity's most important goal. Having a plan to confront global risks could help achieve this goal. Although some effort to prevent various risks has been made, society as a whole continues to be absorbed by its petty internal conflicts.

However, there was a time when the struggle against what was perceived as a global risk was at the peak of international attention and resulted in mass street actions. This is the anti-nuclear struggle in the 80s. (A nuclear war is unlikely to lead to the complete destruction of humanity, but many thought it could.) And it ended with some success: international treaties were signed, which significantly limited the nuclear arsenals, and the Cold War ended.

It is obvious that sooner or later politicians and parties will appear advocating for prevention of global risks. Now in the United States the Transhumanist Party has been created, it stands for radical life extension, for anti-aging and prevention of global risks.

In addition, studying global risks should take shape of a science with all relevant attributes, namely, a scientific journal, an online forum, a series of conferences, a scientific institute, an inter-governmental panel on risk analysis (similar to the panel on global warming).

It should also provide an opportunity for dialogue between advocates of different points of view and not be restricted to a narrow circle of like-minded people referring to each other.

Reactive and Proactive approaches

Reactive: React to most urgent and most visible risks. Risk are ranged by urgency.

Pros: Good timing, visible results, proper resource allocation, investing only in real risks. Good for slow risks, such as a pandemic.

Cons: can't react to fast emergencies (AI, asteroid, collider failure).

Proactive: Envisioning future risks and building a multilevel defence. Risks are ranged by probability.

Pros: Good for coping with new risks, enough time to build defense.

Cons: Misinvestment in fighting with non-real risks, no clear reward, challenge in identifying new risks and discounting mad ideas.

This map is based on the proactive approach, but now the reactive approach to risks is dominating. We can't state that the proactive approach is always better as it may lead to excessive activity in such areas that are best left alone. But a proactive study of future risks is needed.

A1.1-Step 3. International cooperation

Super-UN

•All states contribute to the UN to fight certain global risks

• States cooperate directly without UN

• Superpowers take responsibility for x-risk prevention and sign a treaty

• International law about x-risks is introduced which will punish people for rising risk: underestimating it, plotting it, neglecting it, as well as reward people for lowering x-risks, identifying new risks, and for efforts to prevent them.

• International agencies dedicated to certain risks (old ones such as the WHO and new ones for new risks).

Stimuli

• A smaller catastrophe could help unite humanity (pandemic, small asteroid, local nuclear war)

• Some movement or event that will cause a paradigmatic

change so that humanity may become more existential risk aware.

This item includes more practical steps to prevent global risks. It is assumed that understanding of the nature of risks and the importance of their prevention is already achieved. The next step is international coordination of efforts.

The UN is the most authoritative international organization created to fight the risk of a new war. On the other hand, the UN in its present form is largely discredited and weak bureaucratically.

As a result, the authority to fight global risks may be delegated not to the UN, but to some other organization, or the strongest military and economic power, such as the United States.

Depending on the type of a risk, not all states can participate in its prevention, for example, only one or several large states should unite to protect themselves against the threat of an asteroid. However, dealing with the most serious risks requires cooperation of all economically developed countries of the world, as well as access to the entire territory of the Earth, without exception.

A good example is the fight against the Ebola epidemic in 2014, which could have become a global risk if its exponential growth had not been stopped. However, President Obama chose the right strategy: maximizing the suppression of the epidemic outbreak in the center of its dissemination (while another proposed strategy, that of announcing a total quarantine for the infected countries, would have lead to millions of deaths and the emergence of permanent foci, where Ebola could have evolved into a more dangerous form). Many developed countries and international organizations, such as the WHO, participated in the fight against Ebola.

Although in the first half of 2014 the international community demonstrated extreme laziness and lack of foresight with regard to the exponential process of the Ebola outbreak, later effective mobilization of resources took place. This shows that a small or slow-growing disaster leads to the acceleration of the integration processes and bring different organizations together to solve the problem. However, not all risks will develop so slowly and so clearly.

Practical steps to confront certain risks

Biosecurity:

- Developing better guidance on safe bio-technology
- DNA synthesizers are constantly connected to the Internet and all newly created DNA are checked for dangerous fragments

• Funding controlled environments such as clean-rooms with negative pressure

• Introducing better quarantine laws for travelling during pandemic

• "Develop and stockpile new drugs and vaccines,

monitor biological agents and emerging diseases, and strengthen the capacities of local health systems to respond to pandemics" (Matheny)

Environment

• Capturing methane and CO2, probably, by bacteria

• Investing in biodiversity of the food supply chain, preventing pest spread: 1) better food quarantine law, 2) portable equipment for instant identification of alien inclusions in medium bulks of foodstuffs, and 3) further development of nonchemical ways of sterilization.

• Promoting mild birth control (female education, contraceptives)

• Promoting the use of solar and wind energy

Nukes

• Introducing asteroid deflection technology without nukes

• Improving nuclear diplomacy

• Using instruments in tech to capture radioactive dust

Some practical steps to prevent risks can be taken without a global plan and within individual research programs. Here are listed only some of these steps.

A1.1 Risk control

Technology bans

• Introducing an international ban on dangerous technologies or voluntary relinquishment such as not creating new strains of flu

• Freezing potentially dangerous projects for 30 years

• Lowering international confrontation

• Locking down risky domains beneath piles of bureaucracy, paperwork and safety requirements

Technology speedup

- Differential technological development: develop safety and control technologies first (Bostrom)
- Introducing laws and economic stimulus (Richard Posner, carbon emissions trade)

Surveillance

- International control systems (such as IAEA)
- Internet scanning, monitoring from space.

The logical development of the theme of bans and freezing projects is the idea of differentiated technology development proposed by Bostrom. We must invest in the development of technologies that enhance our security and slow down the development of technologies that increase the risks. As a result, we are not canceling the general trend of progress, but changing the shape of its front. That is, it is necessary to quickly develop technologies that enhance the control and management, namely, AI and surveillance systems, and slow down the development of technologies that can quickly lead to uncontrollable consequences.

So far this idea remains only wishful thinking.

Richard Posner proposed manage risks through legislation and economic incentives. The most famous attempt to do something of the kind has been carbon trading. However, in general we do not see the results of its effectiveness, as carbon emissions and coal burning continued to rise.

Instead of prohibitions, we can allow certain types of activity, but exercise total control over it, so that it would be carried out in a safe manner and in a peaceful way. An example of this kind of activity is controlling the IAEA on nuclear energy. Currently, the technical capabilities to monitor have incredibly widened thanks to cheap electronics, spyware systems in every mobile device, satellite monitoring, scanning the Internet. However, the division of the world into rival states makes such control difficult even for such large facilities as nuclear plants, and the control of individual biological laboratories is even more difficult.

However, the result could be a totalitarian Orwellian society that under the pretext of protection against global and other risks penetrates into people’s private life and makes various abuses. Scandals involving a ubiquitous surveillance occur regularly in the United States. The controllers themselves are out of control and can make the same irregularities that they must prevent.

David Brin offered an alternative society of total control by the state. It is a society of total transparency, where everyone can watch one another through electronic means, and a group of civil society activists scour the world looking for potential terrorists. This idea may be useful but if it is considered the only solution to the problems of global risk, it seems far-fetched.

Elimination of certain risks

- Universal vaccines, UV cleaners
- Asteroid detection (WISE) prove that no dangerous asteroids exist
- Transition to renewable energy, cut emissions, carbon capture
- Stopping LHC, SETI and METI until AI is created
- Getting rogue countries integrated or prevented from having dangerous weapons programs and advanced science

In parallel with the development of prohibitions and means of control, the development of technology can lead to the situation that some risks will get "closed", that is, either means to effectively prevent them will be created, or it will be proved that they are impossible.

For example, the creation of a universal vaccine against influenza, or in general from any viruses, will significantly reduce the risks of biotechnology (such works are going on now, and there are many interesting ideas).

The development of observational astronomy, and, first and foremost, infrared astronomy (space telescope WISE) will reveal all potentially dangerous near-Earth asteroids and most likely prove that none of them is going to threaten the Earth in the next 100 years. This would eliminate the need for the construction of asteroids interception systems, which themselves can be dangerous (because they consist of powerful missiles and nuclear weapons, which could be used as space weapons or against the Earth itself).

The development of solar and wind energy and removing carbon from the atmosphere and using it as a building material will significantly reduce the risk of running out of resources and energy, as well as air pollution and global warming, which in total will reduce the risks of a new world war and increase life expectancy. Of course, it may be not necessary, if we quickly create a strong AI based on nanotech industry, but the timing of this is difficult to predict.

There is also the idea to produce some minerals in space, of which I personally am skeptical, because the differentiation of the Earth's interior and the water cycle produced a significant

enrichment of the primary ores which didn't not happen on other planets and on asteroids.

A1.1 – Step 4: Second level of defense on high-tech level: Worldwide risk prevention authority

- Establishing a center for quick response to any emerging risk: x-risk police
- Introducing worldwide video-surveillance and control
- "The ability when necessary to mobilize a strong global coordinated response to anticipated existential risks" (Bostrom)
- Peaceful unification of the planet based on a system of international treaties
- Robots for emergency liquidation of bio- and nuclear hazards
- Narrow AI based expert system on x-risks, Oracle AI

After some risks have been prevented with the help of specialized agencies and with individual measures, a clear need will emerge for set up an agency responsible for preventing any future global risks. This may be the UN Security Council, or a UN committee vested with adequate powers.

The first of these powers should be worldwide gathering of information on emerging risks with the use of all possible technical means.

Another of its powers would be the ability to rapidly respond to emerging threats, such as sending out troops and medical teams to the location where an epidemic started, or even to carry out nuclear strikes on laboratories that produce harmful biological or nano-replicators.

It would only be possible to create such an agency if there is a peaceful unification of the world into a single supranational structure. Peaceful reunification is possible in the first place through a complex system of agreements, similar to the system that provides the integration of the EU. Probably the subjects of the integration will be supranational entities rather than individual states.

However, in this scenario there is a dangerous divergence which may be called a "war for the unification of the planet."

Planetary unification war

- A war for the world domination

- One country uses bio-weapons to kill all the world's population except its own which is immunized
- A super-technology (such as nanotech) is used to quickly gain global military dominance
- Doomsday Machine blackmail

This scenario I have outlined in red, because it is not good, and could very likely lead to the destruction of all mankind or a significant part of the world's population. That is, this scenario is not desirable but it looks increasingly likely. Namely, instead of the integration and domination of humaniatrian values we see the division of the world into blocks, and there is a group of rogue states not wishing to integrate with any of the blocks (North Korea, Islamic State) and at the same time actively developing weapons of mass destruction.

In principle, any world war is a war for world domination, but no war ever ended up by one country winning it. A future world war may also not have a winner, and only become meaningless homicide and a catalyst for the development of ever more dangerous weapons.

It may happen in the future that there is one winner who has crushed some of his opponents and persuaded the others to obey by threats.

Such a war may be a conventional, or nuclear, or based on supertechnologies. The latter option is most likely to lead to the victory of one party, as supertechnologies can give a decisive advantage.

Even worse scenario is that of one country creating a doomsday weapon it blackmails the rest of humanity with and makes the rest of humanity capitulate to it. But if such weapons are created by a number of countries that have mutually exclusive conditions of deployment of those weapons, then we are doomed (Herman Khan).

Finally, the worst planet unification war scenario is that of one country destroying all the others, e.g., by using a virus against which its own population is vaccinated.

A planet unification war is a very bad method to unite the world, but a method that, unfortunately, may work.

Active shields

- Geoengineering against global warming
- Organizing a worldwide missile defense and an anti-asteroid shield
- Putting up a nano-shield – a distributed control system to

control hazardous replicators

- Putting up a bio-shield – a worldwide immune system
- Establishing dangerous memes control (existential terrorism prevention)
- Controlling the knowledge of mass destruction
- Amalgamating the state, the Internet and the worldwide AI into a worldwide monitoring, security and control system
- Isolating risk sources at a great distance from the Earth;
- Performing scientific experiments in such ways that are close to natural events

So if the unification of the planet, at least in the field of prevention of global risks, happens more or less successfully, it will become possible to implement a number of technical measures to prevent future risks.

At the same time, taking into account the exponential development of technologies, the risk that will be the most dangerous in the middle of 21st century is the risk of replicators (bio, nano or computer virus), and in order to control them we need different types of high-tech shields.

By the mid-century the national state will have integrated with different means of AI and robotic systems. As a result, so-called "active shields" will emerge, a kind of a global immune system capable of detecting certain dangerous replicators or other risks and preventing them instantly, perhaps even without human involvement.

First among them should be named geoengineering, that is the control of the global temperature by means of spraying water into the upper atmosphere or sequestrating carbon dioxide from the atmosphere.

An international missile defense system may also be considered a global shield, although it is unlikely to be necessary if all the countries in the world get united.

A bio-shield will test DNA of various organisms in the environment with the aim to immediately detect dangerous viruses and replicators, in which in is similar to the human immune system.

A nano-shield will appear at a later stage of development, when nanorobots-replicators have been created, and there is the risk that they may start to multiply in the environment, that is, they will become "gray goo”. In order to control that it is necessary to accommodate specialized sensors everywhere in the environment, even in the world's oceans (Robert Freitas).

A system of control over criminal activity and potential terrorism can also be regarded as a global shield, as well as a system of

control over such an artificial intelligence that can start self-improving or planning to perform some destructive activity.

Ultimately, the global AI will control all the shields, incorporate both the Internet and various government agencies take over their functions.

Finally, there is the idea to move some of the sources of risk to a considerable distance from the Earth so that they would not cause harm if something goes wrong or would allow some time for preparation. This applies to dangerous biological experiments and dangerous physical experiments, but is unlikely to deal efficiently with dangerous attempts to create a self-improving AI that could easily "escape" by communication channels.

We do not touch here upon the problem of creating a safe AI, assuming it will be created, if possible, within the framework of the Plan A2. In addition, I have a separate map showing ways to create a safe AI, which is extremely complex, and which contains about a hundred possible ideas.

If a friendly AI has been created it may interrupt implementation of the Plan A1 at any stage and offer better solutions

Step 5 – Reaching indestructibility of civilization with negligible annual probability of global catastrophe: Singleton

- Singleton is “a world order in which there is a single decision-making agency at the highest level” (Bostrom)
- Setting up a worldwide security system based on AI
- Developing a strong global AI preventing all possible risks and providing immortality and happiness to humanity
- Colonization of the solar system, interstellar travel and Dyson spheres
- Colonization of the Galaxy
- Exploring the Universe

The result of the implementation of the Plan A1 and a number of other plans should be the creation of what Bostrom calls Singleton, a single center of decision-making within the civilization that uses AI and ensures prevention from all global risks.

AI is the most effective tool for adaptation, and therefore by definition must be able to prevent all the risks that are generally preventable. In addition, it should solve the problem of “good” for humanity in the broadest sense of these words, including the removal of aging, death, suffering, and also in other aspects that are still difficult to understand for us.

It should ensure further unlimited development of mankind so that the potential of the human species would be fulfilled to the maximum possible extent. Probably, this could be achieved by having combined AI and human beings. After that a protected and immortal humanity will face the task of colonization of the solar system, the Galaxy and the Universe.

Humanity will become a Kardashov 2 and 3 level civilization.

A strong AI and, consequently, Singleton will most likely be created, if that is possible at all, before the end of the 21st century, at the earliest by 2030. So the period of global risks in the history of mankind will last no more than a hundred years, after which humanity will either perish or reach a certain state of indestructibility.

Plan A1.2 – Decentralized risk monitoring

This plan largely originated from the crowdsourcing of ideas based on a previous version of the map, during which more than 20 interesting suggestions appeared.

The essence of this plan is to collect all positive alternatives to the Plan A1.1 and avoid the totalitarian control risks:

- The need for a world war for the unification of the planet,
- An Orwellian worldwide totalitarianism, with its restriction of freedom, the penetration of the state into private life
- And, above all, the risk, built into any totalitarian system, of a failure in the centralized control system. After all, a control center itself is not accountable to anyone, it is out of control. Centralized control has an area which is out of control: that is the top of control pyramid, and even control over control doesn't solve this intrinsic problem.

This version of the Plan A1 is good, positive, but perhaps a little bit naive. In reality, some of its elements may be combined with a control version resulting in a more viable solution.

The steps in this plan are different. It starts with changing human values, proceeds to changing behavior and society and then to the organization of mutual control.

A1.2 – 1.Values transformation

The value of existential risk reduction

- "A moral case can be made that existential risk reduction is strictly more important than any other global public good" (Bostrom)
- Making the value of the indestructibility of civilization the first priority on all levels: in education, on the personal level, and as a goal of every nation

• Improving the public desire for life extension and global security

Dissemination of this value

• Reducing of radical religious (ISIS) or nationalistic values

• Raising the popularity of transhumanism

• Promoting movies, novels and other works of art that honestly depict x-risks and motivate their prevention

• Introducing memorial and awareness days: Earth day, Petrov day, Asteroid day

• Educating in schools on x-risks, safety, and rationality topics; raising the sanity waterline

To a large extent the prevailing policies are determined by the values prevailing in society. If you describe it very crudely, there are two large opposing groups of values: the first group is national-religious values, and the second group is the value of the progress of humanity, unity and life extension.

The national-religious group is characterized by believing in afterlife, following unproven dogmas, the primacy of the value of the group over the value of the individual or humanity as a whole. There are many such groups and they conflict with each other.

For such a group humanity as a whole and its fate are not values, and a global catastrophe could even be desirable as a religious objective. Unfortunately, we see that the popularity of such groups is only growing in all the countries of the world.

I is typical of such groups to be at war with each other (Shiites and Sunnis), and especially at Western values (Boko Haram against education).

In a milder form, these values are represented in Western countries as nationalist and religious movements.

For the second group the life of the individual and the fate of humanity as a whole are important. In general, this group can be called Western values or universal human values.

Its logical conclusion is the philosophy of transhumanism, which declares the absolute value of human life and the need for its indefinite extension, as well as the importance of the prevention of global risks. However, the spread of the transhumanism is very slow. In part because such values coexist with a lot of other values, such as the traditional religion and even the environmental

movement.

Paradoxically, despite the technological advances of the recent decades, national and religious values are experiencing a renaissance.

In parallel to the transformation of values, the transformation of the picture of the world is underway. In fact, each block of values implies a certain view of the world. What "Western" and "traditional" values have in common is the idea of the future being quite linear. If you take a future vision with exponential development, it immediately raises the questions of global risks and human immortality.

Another problem associated with the existence of different values is the existence of nation states with different identities, different government structures different declared values, and every nation state has its own egoistic interests. Their relation with values is too complex to try to elaborate it here, but their existence is a major contribution to existential risks due to possible wars, arm races, terrorism, control prevention and different levels of control in different parts in the world (which means that criminals could find the least controllable place, like Somali).

Arms races may cause dangerous technologies to be developed faster than the methods of their control.

On the other hand, former enemies are able to unite in the face of imminent danger, if it becomes visible.

We could think of changes of values as changes in the probability of different types of events. There will always be people and groups with opposing values but if human life value dominates, it would mean less violence (an we see a decrease in violence over centuries). If the value of a future generation is high most people will be less likely involved in activities raising chances of global risks.

So the value effect is indirect and hard to measure but it could change extinction probability by the order of magnitude.

Ideological payload of new technologies

The idea is to design a new monopoly tech with a special ideological payload aimed at global risks prevention.

Any new tech suggests a new norm of behavior. Here are listed new technologies and values that they promote.

- Space tech – Mars as backup, long term survival
- Electric car – sustainability
- Self-driving cars – risks of AI and value of human life
- Facebook – empathy and mutual control
- Open source – transparency
- Computer games and brain stimulation – virtual world

A1.2 – 2: Improving human intelligence and morality

Intelligence

- Nootropics, brain stimulation, and gene therapy for higher IQ
- New rationality: Bayesian probability theory, interest in long term future, LessWrong
- Fighting cognitive biases
- Many rational, positive, and cooperative people are needed to reduce x-risks (effective altruists)

Empathy

- High empathy for new geniuses is needed to prevent them from becoming superterrorists
- Lower proportion of destructive beliefs, risky behaviour, and selfishness
- Engineered enlightenment: use of brain science to

make people more united, less aggressive; opening the realm of spiritual world to everybody

Morality

- Preventing worst forms of capitalism: the desire for short term monetary reward
- Promoting best moral qualities: honesty, care, non-violence

The idea here is that if people become better, then the probability of accidents will decrease. More intelligent, more moral, more responsible people are less likely to be ill-intentioned or commit a fatal error.

Intelligence (IQ) correlates with less violence and a longer life, it helps predict consequences of one's actions.

Empathy will lower violence and help adapt a holistic world view and the value of preservation of human civilization.

Morality will make people act less violently and more altruistically.

A1.2 – 3. Cold War, local nuclear wars and WW3 prevention

- Establishing an international conflict management authority — an international court or a secret institution
- Implementing a large project that could unite humanity, such as a pandemic prevention project
- Integrating rogue countries into the global system based on dialogue and appreciation for their values
- Introducing hotlines between nuclear states
- Promoting antiwar and antinuclear movement
- Using international law as the best instrument of conflict solving
- Peaceful integration of national states
- Employing cooperative decision theory in international politics ("do not press on red")
- Preventing brinkmanship
- Preventing nuclear proliferation

Dramatic social changes

- These could include many exciting but different topics: a demise of capitalism, a hipster revolution, internet connectivity, global village, dissolving of national states.
- Changing the way politics works so that the policies implemented actually have empirical backing based on what we know about systems.
- Introducing a world democracy based on Internet voting.
- Maintaining high-level horizontal connectivity between people

A1.2 – 4. Decentralized risk monitoring

- **Transparent society**: everybody can monitor everybody:

- groups of vigilantes scanning the open Web and sensors
- "Anonymous" style hacker groups: search in encrypted spaces

- **Decentralized control**:

- local police handle local crime and terrorists;
- local health authorities identify and prevent the spread of disease
- mutual control in professional space
- Google search control
- whistle-blowers inform the public about risks and dangerous activities

• **Net-based safety solutions:**
- ring of x-risk prevention organizations
- personal safety instructions for every worker: short and clear

• **Economic stimuli:**
- carbon emissions trade
- prizes for any risk identified and prevented

• **Monitoring of smoke, not fire:**
- search predictors of dangerous activity using narrow AI

Plan A2. Creating Friendly AI

A2.1 Study and Promotion

- Study of Friendly AI theory
- Promotion of Friendly AI (Bostrom and Yudkowsky)
- Fundraising (MIRI)
- Slowing other AI projects (recruiting scientists)
- FAI free education, starter packages in FAI

The basic idea, the terminology and the development issues related to a friendly AI are defined by E. Yudkowsky and Nick Bostrom. Yudkowsky created MIRI and LessWrong.

The basic idea is that a strong, self-reinforcing Artificial Intelligence is a global risk, but if you make it "friendly" it is going to be safe for the people and be able to prevent all other global risks, as well as to solve other problems of mankind, and, moreover, will be a source of a huge number of various benefits that we are unable to specify at present, but which will include prevention of aging, suffering, involuntary death and creation of much happier human lives.

However, we do not know how to create a human-level AI, and. moreover, do not know how to make it friendly. Because of this, we need at the beginning to conduct a thorough study on the methods of implementation friendliness, and collect a team of scientists and the money to do it.

Gradually, it is starting to happen. Bostrom's book "Superintelligence" basically retells the ideas previously expressed by Yudkowsky, but Bostrom has been much more successful as an academic and also received the support of Elon Musk who allocated $10 million in 2015 in grants to study ways to create a safe AI. In recent years, articles about the risks of a strong AI have been published in many respected media, and the topic has become widely known.

However, so far the situation is that there are about a hundred ideas on how we can create a safe AI, but there is not one that would look bulletproof. Thus, long-term studies are needed.

But the development of AI is going very fast, which can be seen in the example of image recognition systems and self-driving cars. It is possible that a strong AI will be created by 2030, as was proposed by Vinge in 1993.

One way to create a friendly AI is to increase the number of scientists working on its development as well as improving the overall rationality in society.

Another way is slowing down the development of the whole AI industry, which, for example, may come about through pumping brain out of it or as a result of economic recession. This, of course, will not work.

A2 – 2. Solid Friendly AI theory

- Theory of human values and decision theory
- A full list of possible ways to create FAI, and a sublist of best ideas
- An AI that is proven safe, fail-safe, intrinsically safe
- Preservation of the value system during AI self- improvement
- A clear theory that is practical to implement

The next step should be the creation of the theory of a friendly AI. It should include a number of blocks, such as the theory of value and the theory of decision-making. This theory must mathematically prove that the AI will be safe.

This theory also should be easy to apply in practice. That is, it should be simple to understand, applicable to different AI architectures, should be convincing, and may consist of several independent units. It should also provide multi-layered protection.

Also, AI's self-improvement must not affect its system goals.

I have a map of different ways to achieve AI safety.http://lesswrong.com/lw/mid/agi_safety_solutions_map/

A2.3 AI practical studies

- Narrow AI
- Human emulations
- Value loading
- FAI theory promotion to all AI developers; their agreement to implement it and adapt it to their systems
- Tests of FAI theory on non self-improving models

It is not enough to develop a good theory of Friendly AI, it is also important that it will be applied by a team that will first come close to the creation of a strong AI. In order for the latter to happen, we need to present the theory to all teams, including Google, IBM, Facebook, start-ups in the field of Deep learning, state and military secret projects, as well as foreign companies and individuals such as (possibly) AI hackers.

Another approach is that the same team that created the friendliness theory, should create a friendly AI using its intellectual advantage, but it is unlikely as it is too complex a task.

In addition, the theory will require some adaptation for a specific method of creating AI, or that method must be adjusted to the theory. Then it can be tested on a toy model of AI, somehow kept from self-improving.

A parallel research in human brain may result in its emulations, the creating of specialized AI systems, as well as a research in goals loading into rational agents. All of the above should help us better understand how friendly AI theory should work.

Seed AI

Creation of a small AI capable of recursive self-improvement and based on Friendly AI theory.

Superintelligent AI

- Seed AI quickly improves itself and undergoes "hard takeoff"
- It becomes a dominant force on the Earth

- AI eliminates suffering, involuntary death, and existential risks
- AI Nanny – creating a super AI that only acts to prevent other existential risks (Ben Goertzel)

In fact, there are two points of view on the development of a strong AI. One is that it will be created in a small private laboratory thanks to a single design breakthrough, when it starts quickly self-improving then easily steals away from the laboratory into the Internet, and then takes over the world with good or bad purposes.

The other extreme view is that the AI will be created by the military or some intelligence agency or a large state or semi-state company that has unlimited funding for the purchase of computers and human brain. And this company will also have 10+ years of theoretical advantage (e.g., due to secret mathematical theorems used in encryption) over open sources of information. The AI will self-improve quite slowly (not over hours but over years) and its access to the Internet and other external networks will be opened by its creators intentionally. There could be many intermediate solutions.

What is important is that the principal outcome of this development will be the same: an AI controlling the world, with a certain goal system. In this map we do not touch upon the complexities of creating a friendly AI and many ways in which an attempt to create it can fail, to which I devote two separate maps.

The self-improvement process of a seed AI is risky, and it can result in a global catastrophe that will destroy humanity.

If successful, we get the same Singleton as in Plan A1, so these paths converge. We could also think of Plan A1 as a story about state gradually converting into an AI.

Which of these plans is better? Plan A1 is more suitable for the prevention of global risks that arose prior to the creation of a strong AI. Plan A2 depends on whether AI is possible at all and weather we are able to control it. That is, at the start Plan A1 is stronger, but Plan A2 is stronger at later stages. Therefore, they can be implemented in parallel.

Although if there are several competing systems of AI, it will lead to a war between them and an ensuing disaster.

UnfriendlyAI

- Kills all people and maximizes non-human values (paperclip maximiser)
- People are alive but suffer extensively

Another possible outcome is creation of a strong unfriendly AI that would destroy humanity one way or another. However, in a

certain sense it may better than other ways of extinction as even the most unfriendly AI will carry the knowledge of mankind, and perhaps will be able to create or simulate people, for example, to assess the incidence of other AIs in the Universe. And it may be better than oblivion. Or maybe not, if simulated people will suffer and be doomed.

Plan A3. Improving Resilience

The idea here is that if we increase the resilience of infrastructure and people to any source of death, and do it faster than new means of destruction are developed, humanity will be immune to any catastrophe.

Briefly, the slogan of this plan is to "become immortal".

The plan as a whole is more complicated and less likely to succeed than plans A1 and A2, but some of its elements can be implemented in parallel.

A3 – 1.Improving sustainability of civilization

- Implementing intrinsically safe critical systems
- Promoting a growth in the diversity of human beings and habitats
- Employing universal methods of catastrophe prevention (resistant structures, strong medicine)
- Building reserves (food stocks, seeds, minerals, energy, machinery, knowledge)
- Establishing a widely distributed civil defense, including:

- temporary shelters,
- air and water cleaning systems,
- radiation meters, gas masks,
- medical kits
- mass education

Firstly, additional layers of security should be introduced in all hazardous systems. This applies to control systems of reactors, aircrafts, nuclear and biological laboratories.

Secondly, it should be noted that people are quite homogeneous genetically because our population has recently, about 70 000 years ago, passed through a bottleneck. Between any two chimps there is more difference than between any two human beings. This makes mankind especially vulnerable to infections, the usual protection from which is genetic diversity. As a result of experiments in the

creation of post-human hybrids, chimeras, genetic editing of human DNA can create a new subspecies of man resistant to possible artificial epidemics.

Finally, universal preventive means, such as a universal vaccine. are instrumental to counter entire classes of risks.

German Khan wrote that the best way to win a nuclear war is to possess a high-quality civil defense capable of going through enemy retaliation. The strengthening of preventive means includes developing emergency medicine and vaccine production technologies, and analyzing the samples of the pathogens.

A3 – 2. Useful ideas to limit the scale of catastrophe

• Limiting the impact of a catastrophe by implementing measures to slow down the growth of areas impacted:

- using technical instruments for implementing quarantine,

- improving the capacity for rapid production of vaccines in response to emerging threats

- growing stockpiles of important vaccines

• Increasing preparation time by improving monitoring and early detection technologies:

- supporting general research into the magnitude of biosecurity risks and opportunities to reduce them

- improving and interconnecting disease surveillance systems so that novel threats can be detected and responded to more quickly

• Worldwide x-risk prevention exercises

• Ensuring the ability to quickly adapt to new risks and envision them in advance

A3.3 High-speed Tech Development needed to quickly pass risk window

• Investing in super-technologies (nanotech, biotech, Friendly AI)

• High speed technical progress helps to overcome slow process of resource depletion

• Investing more in defensive technologies than in offensive technologies

The period of global risks is a historical period, relatively speaking, from the creation of nuclear weapons to the creation of a strong AI. It is a kind of adolescence for a civilization, when it can do everything, but still cannot quite control itself. This period will last approximately 100 + / - 50 years.

There is the idea to rush through this period more quickly.

This is partly due to the fact that while some of the risks increase exponentially within the period (biotech & AI), other risks are linearly distributed therein. By accelerating technological progress, we can accelerate exponential risks, as we have less time left to think about how to control them, but reduce linear risks, since the whole period is shorter. In addition, thereby we can reduce the chance of "black swans", which probably are relatively evenly distributed.

Furthermore, if we do not jump on new technologies, we will face challenges posed by older technologies, i.e., we can fall into the trap of Malthusian resource exhaustion. That is, if we do not switch go to new energy sources and production methods, we will in a few decades find ourselves running out of recourses and amidst a civilizational crisis with likely global wars.

A3.4. Timely achievement of immortality on highest possible level

- Researching a nanotech-based immortal body
- Diversification of humanity into several successor species capable of living in space
- Mind uploading
- Integration with AI

This option becomes relevant when with the rapid development of technology, we will be able to upgrade the human body. If we replace all the cells of the body with nanomachines, then no biological infection will be able to do anything with it. Such a body can withstand radiation, cold, live in outer space. Such a person should be afraid only of other nanomachinery or AI (or a virus) taking control of micro robots inside his body.

The logical step beyond that would be transferring human consciousness into a computer, causing it to be able to live in any environment where computers can exist, and it will depend only on AI.

AI based on uploading of its creator

- Friendly to the value system of its creator

- Its values consistently evolve during its self-improvement

The end result of such a race against threatening environmental technologies will be that the person turns into artificial intelligence. Perhaps it will be one person that is the creator of this AI.

Plan A4. Space Colonization

Elon Musk is one of the few who advocate more than one ways to prevent global risks. Most people tend to get hung up on just one.

Namely, Musk speaks about the importance of creating a friendly AI and the importance of moving in space as a means of protection against disasters on Earth. The same idea was expressed by Hawking and many others.

Unfortunately, this idea is weaker than the previous ones, as the space colonization requires the development of space technologies. And these new technologies can also create a disaster that can propagate in space. For example, space rockets can be kinetic weapons. AI can be spread by the radio. Nanobots can fly like dust from one celestial body to another. Biological infection can spread to a spacecraft as well as, for example, AIDS can be carried by people inside ships and aircraft on the Earth. The development of new energy sources can be used for huge explosions in space that can sterilize entire planets and even the Solar System. There could also be a war between space colonies or terrorists inside the colonies infiltrated by hostile propaganda.

So moving to space is not a panacea, and the development of appropriate technologies may even have a negative value. Settling in space will save us only from the weakest of risks, such as asteroid or global warming, with which we can cope even without it. However, space colonization can still increase our chances of survival, especially if we will be able to travel to other stars.

A4.1. Temporary asylums in space

- Space stations as temporary asylums (ISS)
- Cheap and safe launch systems

In this section we will consider tech that is already here or can be created on the basis of the existing technologies in the next 10-15 years.

The International space station (ISS) already exists, but it cannot operate autonomously for more than a year. If mankind dies

very quickly with the environment preserved, the six men and women on the ISS can be the beginning of a new humanity. But the chances of such a disaster are rather small.

In the next few years we can build a base on the Moon, and in a couple of decades a base on Mars. If only 10-15 people get to live on the Moon, the Moon's value as a back drive for mankind will be no more than that of ISS.

A4.2. Space colonies near the Earth

- Creation of space colonies on the Moon and Mars (Elon Musk) with a population of 100-1000 people.

Elon Musk is now building a space launcher that could deliver 100 people on Mars and it could fly in 2020s.

A 1000-people colony on one of the nearest celestial bodies can exist independently for decades but it still will not be self-sufficient or able to continue technological progress. And it, possibly, represents the upper limit of what we could reach on our current space tech level.

If a million people lives on Mars, then they will probably be capable of self-sustaining, and become the basis for a second humanity, without even coming back to the Earth, if it is lost. With the current technologies sending so many people to Mars will take several decades and huge amounts of money that could be spent in a different way in the world. That is, there is an opportunity cost.

Colonization of the Solar System

- Setting up self-sustaining colonies on Mars and large asteroids
- Terraforming planets and asteroids using self-replicating robots and building space colonies there
- Setting up millions of independent colonies inside asteroids and comet bodies in the Oort cloud

This option involves the colonization of the Solar System on the basis of next generation technology, robots and robots-replicators. Such a colonization may be much easier and cheaper: in principle, only one robot-replicator could start a Solar System wave of colonization and it could be build by a private person.

However, there is more risk involved: nanorobots can get out of control, or be used to create dangerous giant structures in the Solar System, or fight each other, or simply become space-gray goo.

A4.3. Interstellar travel

- “Orion” style, nuclear powered “generation ships” with colonists
- Starships operating on new physical principles with immortal people on board
- Von Neumann self-replicating probes with human embryos

The first idea is that of a generation starship traveling rather slowly and with people living and having children on board. Such a starship can be built on the basis of modern technology with a budget of about a trillion dollars. That is the project "Orion" envisioned in the 60s, a spaceship driven by explosions of nuclear bombs. It is quite feasible, although rather cumbersome and not environmentally friendly. It can reach the nearest star in 40 years.

Another idea that involves using space travel as means of escaping global risks is to move through space so fast that no local impact will be able to influence the whole population of human civilization. (It also means communication channels are off.) So it was in human past, when the means of transport were very slow (ships, vehicles). This requires interstellar travel with near-light speed.

Of course, there is a chance that there will be new spaceships available built on new physical principles. But with new principles new risks will come forth. Even if the Orion Spacecraft explodes at the start or turns into a kinetic weapon, it could threaten life on the Earth. New principles of space travel would mean new sources of energy and new ways of spreading damaging effects which is a recipe for new global risks.

One more option is using von Neumann probes, that is, interstellar robots-replicators. They can be loaded with human embryos (or DNAs), which will be brought up by a robot-nanny. The mass of such a starship could be only a few grams.

This item in the map is connected by a vertical yellow stripe to another block (creating nanotech immortal bodies) which means a strong connection. Such nanotech bodies will probably be able to live in space.

Interstellar distributed humanity

- Many unconnected human civilizations
- New types of space risks (space wars, planets and stellar explosions, AI and nanoreplicators, ET civilizations)

As a result of space colonization, we may get something like a loosely bound "galactic empire". Some of the planes will die, others will fight wars with each other, and others will thrive.

Plan B. Survive the catastrophe

The best way to escape global risks is to prevent a global catastrophe. The higher the technological level when it happens, the harder the catastrophe will be to survive. On the other hand, there is a scenario, which I call "the oscillations before the singularity," when a very large catastrophe precedes the creation of truly strong supertechnologies. For example, the proliferation of low-cost nuclear weapons will lead to an intense nuclear war that in turn will result in humanity regressing to an earlier stage. Or a pandemic destroying 90 percent of the population.

For a variety of reasons the chances of such a "semi-global" catastrophe are two to three times higher than that of a global catastrophe. (Like Pareto's law of the distribution of risks: for example, one dead per two or three wounded in different accidents.)

In this case, the availability of shelters can play a key role in the survival of humanity.

On the other hand, the hope for survival in an asylum should not be very high.

If shelters are super complex and expensive, they will be too scarce and can become targets in a nuclear war.

If they are numerous and cheap, they cannot be good enough to provide full long-term autonomy.

In addition, no asylum types are universal. Every asylum is designed to provide protection against a certain type of disaster. For example, before World War II in the USSR they built shelters against chemical weapons with thin walls, which proved to be completely useless against conventional bombing.

Inside a shelter, you can continue to produce weapons or dangerous viruses. Shelters cannot protect from dangerous nanorobots and AI, and certain types of biological weapons, such as one that spreads slowly and secretly.

Therefore, it is better to invest in different versions of Plan A, and do not to hope for an asylum. But a certain number of shelters would not hurt.

B1. Preparation

- Fundraising and promotion
- Writing a textbook on rebuilding civilization (Dartnell’s book “Knowledge”)
- Stockpiling knowledge, seeds and raw materials (Doomsday vault in Norway)
- Founding survivalist communities

The first step in creating shelters should be creating the project and allocating the money. Most likely, it will be done not by individuals but by states that build shelters in case of a nuclear war.

A good idea is to create a knowledge bank to restore civilization from scratch and mount it on heavy-duty vehicles. Dartell's book "Knowledge" is such an attempt.

A "doomsday vault" is a storage of seeds in Norway which may be used after a large-scale catastrophe and as a matter of fact it is used to store knowledge.

Also, the survivalists movement, whose members mainly train for fun to survive in difficult conditions, could become useful in the case of certain types of global catastrophe.

B1 is connected to A4.1 ("temporary asylum space"), since in fact they are the same thing.

B2. Buildings

- Building underground bunkers, space colonies
- Seasteading
- Converting existing nuclear submarines into refuges

Natural refuges

- Uncontacted tribes
- Remote villages
- Remote islands and Antarctica
- Oceanic ships
- Mines

The ideas to construct autonomous ultra deep bunkers sound pretty crazy, but bunkers at depths of up to 1 km with autonomy about 1 year are quite real and probably already exist. Mines may be converted to them.

Nuclear submarines are designed for long autonomous existence. Their autonomy is about one year.

Distant isolated tropical islands can serve as shelters in the event of a pandemic. Some islands completely avoided the Spanish flu.

Deserted villages in the forest can also make good refuges. Tribes that never had contact with the outside world may well survive humanity.

"Water World". The sea is full of ships. Many of them have a high degree of autonomy, for example, can go for long fishing trips, some have standalone nuclear engines. Tankers carry a large amount of fuel, and container ships are full of food and goods. They can survive some accidents, especially a nuclear war or a biological attack.

There is also the Seasteading movement aimed at the creation of autonomous communities floating above the sea. These settlements can also withstand some types of disasters.

Another type of buildings that can be used as human refuge are research stations in Antarctica. They also have great insulation and autonomy. Deep mines could also help survive miners working inside.

B3. Readiness

- Crew training
- Distributing crews to bunkers
- Implementing crew rotation
- Building different types of asylums
- Freezing embryos

It is not sufficient to have a refuge: you also need to prepare people for living in it. Pre-built shelters should be crewed with well-trained and healthy men and women who must have the skills to build a civilization from scratch, and fitted with appropriate instruments. It is necessary to carry out regular crew rotation: one team is resting while the other is "waiting for disaster." Accidentally caught in a bomb shelter people can be extremely ineffective in restoring civilization.

We can also use the power of the law of large numbers, that is to have a lot of very different shelters in different parts of the world hoping that this may work out fine?

There is still the idea of robotic shelters with frozen human embryos somewhere in the ice of the Antarctic. While it is now impossible, it may be real in 20 years from now. If sensors stop receiving signals of the existence of the terrestrial civilization, the system will wait for 10 years and then get activated and start an artificial uterus, people will be born, they will be taught by robots, and humanity will be restored. This, of course, is not so simple.

B4. Miniaturization for survival and invincibility

- Building adaptive bunkers based on nanotech
- Colonizing the Earth’s crust by miniaturized nano-tech bodies
- Moving into simulation worlds inside small self-powered computer systems

This scenario is close to science fiction but still needs to be mentioned. Future tech will allow the creation of advanced protection systems much more sophisticated than simple

underground buildings.

This presents a problem of the sword and the shield. If a shield is always stronger, it means that a catastrophe is preventable at any level of tech development. But this also means that the shield should be based on the same tech level as the sword or even be more advanced.

B5. Rebuilding civilization after catastrophe

- Rebuilding the population
- Rebuilding science and technology
- Preventing future catastrophes

What we need is not only to survive a disaster but to be able to rebuild our civilization, human population and technology. Moreover, we need to learn lessons from past disasters and prevent future disasters.

Restarting civilization from scratch is not easy since most of the easily accessible deposits of resources will have been exhausted. Using the ruins of human civilization as a source of scrap metal will not contribute to the development of a sustainable self-sufficient society.

According to scientists, a human community capable of self-renewal should include about 1,000 members, smaller communities will be threatened by degradation and destruction due to accidental coincidence of circumstances (R. Hanson).

Reboot of civilization

- Several reboots may occur
- In the end, there will be a total collapse or a new level of civilization

The success of the shelters strategy is that survivors will restart civilization in its entirety, which may require several hundred or even thousands of years. If semi-global disasters occur rather frequently, it may take several cycles to restart. Ultimately, however, there are only two stable states: ultimate destruction or transformation into super-civilizations, immune to global risks. (link)

Plan C. Leave Backups

This plan is, in a sense, a gesture of despair. The chances of its success are very small, and that success is quite illusory.

The idea is that we are not the last civilization in the Universe, and someone will find our remains and will resurrect us using them. It will be either a next terrestrial civilization, if life on the Earth survives, or an extraterrestrial civilization, if they exist and are capable of interstellar travel.

C1. Time capsules with information

- Underground storages of information and DNA for future non-human civilizations
- Eternal disks from Long Now Foundation (or *M-disks*)

The idea is to leave information on media that can exist tens of thousands or even millions of years. The task of creating them is not quite simple, but DNA samples can remain intact for such a long period of time. "M-disk" format is designed for 1,000 years of storage. Long Now Foundation is developing a device that can work and store information for 10 000 years. DNA strains are recoverable for millions of years and maybe even longer if preserved in a cold place.

C2. Messages to ET civilizations

- Sending out interstellar radio messages with encoded human DNA
- Creating storages on the Moon, other planets
- Storing frozen brains in cold places
- Sending out Voyager-style spacecrafts carrying information about humanity

METI, or sending a radio message into space, can also serve as a means to save the information (although it carries risk of attracting the attention of dangerous extraterrestrial civilizations).https://en.wikipedia.org/wiki/Active_SETI

Until now very few transmissions occured and the chances that they will be received by someone are negligibly small. Radio and television programs broadcast from the Earth carry a lot of information but they will probably dissipate in space before they reach any possible civilization.

We can also create a storage on the Moon, perhaps on a pole or in a cave, where the eternal cold and the lack of geological changes and radiation will preserve its content for tens of millions of years, and perhaps longer. On the Moon we would store digital

information, artifacts, tissue samples and even plasticized or frozen human brains as well as DNA.

In addition, there are several remains of the spacecraft that once were sent to Mars, other planets and out of the Solar System. Some of them contain brief messages to aliens engraved on metal plates. https://en.wikipedia.org/wiki/Voyager_Golden_Record

C3. Preservation of earthly life

- Creating conditions for re-emergence of new intelligent life on the Earth
- Directed panspermia (Mars, Europe, space dust)
- Preservation of biodiversity and highly developed animals (apes, habitats)

A new civilization could arise on Earth after humans and we should strive to preserve the species that most likely would give rise to it, that is most highly developed mammals (i.e., monkeys, dogs, rats, dolphins) as well as birds. Some chimps already are using tools http://news.discovery.com/animals/female-chimps-seen-making-wielding-spears-150414.htm and probably could develop general intelligence in several million years. Other species it would take tens of millions years and we should bear in mind that due to rising solar luminosity all earthly life will go extinct in 100 million to 1 billion years, and it is complex life that will die off first http://www.sciencedaily.com/releases/2013/12/131216142310.htm . It could also happen much earlier if various positive feedback about global warming is taken into account, including CO2, methane and water vapor as greenhouse gas.

The higher developed life survives, the faster it may form a new intelligent civilization that can then finds traces of humanity.

Microorganisms will survive almost any kind of disaster as they exist at depths of up to several kilometers, but hardly one of them can develop new multicellular life, because it may take about a billion years and before that the heating of the Sun will have made the Earth unsuitable for life.

It is important to preserve the integrity of the biosphere because only as a whole it will be able to evolve and give rise to new intelligent life.

We can also spread life beyond the Earth, which is called directed panspermia.https://en.wikipedia.org/wiki/Directed_panspermia Mars and Jupiter's moon Europa are best suited for this purpose in the Solar System. About ten dwarf planets and moons in the Solar System have under-ice oceans of liquid water, and they could be fertilized by earthly life (although we should check beforehand if

there is any local life). We can go even further and send some dust with frozen microorganisms in the direction of the nearest stars.

If life gets spread across the Galaxy, then sooner or later it will find a new planet, which may result in a complex biosphere and intelligent life. This intelligence can then return to the Solar System and find traces of humanity, although very few of them will remain after billions of years.

C4. Robot-replicators in space

- Mechanical life: nanobots ecosystem and von Neumann probes based on nanobots
- Preservation of information about humanity for billions of years in replicators
- Safe narrow AI regulating such robots-replicators

It may happen that humanity will perish but some form of mechanical life will remain: some robots-replicators with limited AI. For example, if gray goo appears, then an ecosystem of nanomachines will be formed inside which can store some information or traces of its constructors (an example of this is well described in the novel *The Invincible* by Stanislaw Lemhttps://en.wikipedia.org/wiki/The_Invincible).

If such robots-replicators spread in space, they will be much more stable data carriers than any hard objects or hoards, and can remain relatively unchanged for billions of years. In mechanical systems quite a powerful error correction system could be build, which will prevent their Darwinian evolution and loss of information.

Of course, such devices must be operated by a certain computer program which may be relatively primitive or a narrow AI system unable to self-improve but having superior abilities in some domains, e.g., having the ability to design mechanisms or to adapt to the environment.

Resurrection by another civilization

- Aliens create a civilization that has a lot of common values and traits with the human civilization
- Resurrection of people based on the information about their personalities

This plan can be regarded as successful if the terrestrial civilization, the Homo Sapience specie or personalities of some individuals get brought back to life.

Perhaps it will simply be accomplished by another civilization, similar in some aspects to humans and sharing a significant number of our values and traits.

Plan D. Improbable Ideas

Plan D does not require anything to do but it reflects the hope that something improbable and miraculous will save us. Chances of that, frankly speaking, are small.

D1. Saved by non-human intelligence

- Maybe extraterrestrials are looking out for us and will save us
- We send radio messages into space asking for help if a catastrophe is inevitable
- Maybe we live in a simulation and the simulators will save us

Aside from AI, there are three hypothetical types if supermind that could save mankind:

- aliens or rather an alien AI
- the hosts of the simulation, if we live in that simulation,
- and God.

In any case, we can somehow appeal to the higher mind asking for help and protection, or hope that he on its own should see our problems and save us.

Calling for help to the aliens seems to be the most rational but also the most hopeless option. The difference with Plan C here is that we are not passively leaving traces, but actively demanding help in near-term future, which requires that aliens are very near, that is, they are already hidden in the Solar System or live on the nearest stars, which is very improbable.

So it is very unlikely that ETs exist in the immediate vicinity of the Earth and that we can accurately aim a radio message at them, and that they will have time to arrive before we die, and that their intentions are positive.

Of course, there is a chance that we live in a sort of a cosmic zoo where we are being constantly monitored, and when we achieve a dangerous level, the threat will be eliminated by the help from outside.

But it is also possible to imagine a scenario where space "berserkers" are watching us and will destroy human civilization if it overcomes some unknown to us threshold in its technological development (this may be the development of supertechnologies, nanotech or AI), after which our civilization will become invulnerable for the "berserker". As a result, the chances of

hypothetical benefits from the help of extraterrestrial intelligence are compensated by its hypothetical harm.

The hypothesis of the existence of God can be rationally reduced to the idea that we live in a simulation run by a very high-level and highly moral intelligence. There will be no practical difference for simulation dwellers. Unfortunately, the amount of suffering in the world says that this hypothesis is unlikely (God is or immoral or non-existent).

D3. Quantum immortality

- If the many-worlds interpretation of QM is true, an observer can survive any sort of death including any global catastrophe (Moravec, Tegmark)
- It may be possible to make an almost univocal correspondence between the observer's survival and the survival of a group of people (e.g. if all of them are aboard a submarine)
- Other human civilizations must exist in the actually infinite Universe.

Quantum immortality involves the survival of the observer in at least one line of the possible future. And if even one person is alive, humanity technically is still alive too. Moreover, since those lines of the future, in which only one observer survives and continues to live indefinitely, are less likely than the lines where the group survives, the better chance of surviving is for the group.

On the other hand, it is almost impossible to prove the efficiency of this method, except by surviving for thousands of years due to an incredible combination of circumstances. At the same time, as well as for personal immortality, the largest share of probability can go to sub-optimal outcomes. For example, a group of people will survive as guinea pigs for a hostile AI.

And of course, it is a very hypothetical theory with known objections. Firstly, not all agree with the many-world interpretation of quantum mechanics. Secondly, it is necessary to solve the problem of personal identity.

On the other hand, there is a number of considerations which can enhance quantum immortality if it works, that is, they can increase the share of positive outcomes among the common set of options in which I will survive. For example, if I explicitly associate my survival to the survival of a group of people. For example, if we are all in a submarine. In most cases the destruction of the submarine entails the death of the entire crew. So if I survive, the crew is likely to survive too.

Finally, all this can work even without the quantum theory if we assume that the Universe is infinite or at least very large. In this case, in such a universe there is an infinite number of other civilizations, some of which are very similar to the human one, and

the larger the universe, the smaller the difference, up to an exact match to the last atom. (Tegmark made precise calculations of what the size of the Universe should be to provide the required level of similarity). There is a number of physical mechanisms that could provide the desired size of the Universe, such as the cosmological inflation.

That is, sooner or later humans will arise again somewhere, and some aliens may turn out to be similar to humans as evolutionary mechanisms more or less the same (there is such a term as "convergent evolution", that is the formation of the same form as a result of different evolutionary branches, such as fish and dolphins).

D2. Strange strategy to escape Fermi paradox

A random strategy may help us to escape some dangers that killed all previous civilizations in space

The emptiness of space raises the chances of the conjecture that all previous civilizations have perished, therefore, the civilizational path that seemed rational or conventional, does not lead to safety, and we should choose a random and unexpected path.

If all the previous civilizations have perished, all obvious ways of development may lead to extinction. A strong AI, worldwide totalitarianism or space colonization will not help. If we want to make sure our way of development is different, we need to select it randomly.

But that does not mean that we should let things take their course. That is, in the beginning there should be some kind of global power that will be able to choose a random and unique way.

But the potential harm from this randomness may outweigh the benefit from the choice of the most appropriate strategy. Not to mention the fact that other civilizations can also have used the random approach, and it didn't help, since we do not see them.

Also creating a global power triggers most of global catastrophic risks before it can coherently apply this idea.

D4. Technological precognition

- Predicting the future based on advanced quantum technology and avoiding dangerous world-lines
- Looking for potential terrorists using new scanning technologies
- Creating a special AI to predict and prevent new x-risks

If we could perfectly predict the future in a multiverse, we, probably, could easily avoid global risks. (The knowledge of the inevitable option does not work, cf. ancient tragedies such as *Oedipus*.)

But the strengthening of prognostic tools, the development of futurology, and finally, the creation of artificial intelligence will provide us with a tremendous ability to improve our prediction of the future.

Some new physical effects directly receiving information from the future may help. That is, we need to create a kind of a "quantum radar." For now it remains in the realm of fantasy.

There are also cases where people claim to see prophetic dreams or anticipate the future in another way. Rather, it is a statistical aberration (always something will coincide with something, and our brains are wired to detect coincidences). Perhaps, however, it's worth taking a closer look at the analysis of brain activity in altered states of consciousness.

D5. Manipulation of the extinction probability using Doomsday argument

- Taking the decision to create more observers in case unfavorable event X starts to happen, thereby lowering its probability (method UN++ by Bostrom)
- Lowering the birth density to get more time for civilization

This method is even more esoteric than the previous ones, since it not only uses an unproven mathematical theory, but also uses a clever way to manipulate this theory to influence the probability of future events. This idea comes from Nick Bostrom's article "UN++".

DA is based on the Copernican mediocrity principle: we are most likely in the middle of a group, from which we are randomly selected.

On the one hand, it allows us to predict, for example, the future number of people on the Earth knowing their past numbers, and thus to predict the duration of the existence of human civilization. This is the essence of the classic Doomsday argument.

On the other hand, on the condition that the total number of people that will ever be born in the future with some random events, we could change its perceived probability.

It can be used as described by Bostrom in the thought experiment UN++. In the hypothetical future the UN controls the gamma-ray burst probability, which could significantly harm humanity, by deciding to sharply increase the number of people after this gamma-ray burst. Since this population surge is unlikely

according to the original Doomsday argument, it reduces the risks of this gamma-ray burst.

Unfortunately, this probability shift can be used against anything except the very human extinction.

But there is one idea how to do it. This is the idea to control the number of births per year. The classical DA predicts only the total number of future people which will be about 100 billion, the same as in the past, but due to the high birth rate (about 100 million per year) and a growing world population, the following 100 billion people could be born very quickly, within a few centuries.

However, if you accidentally or intentionally make the birth rate (but not mortality) fall sharply, the next 100 billion people will be born over a very long period.

Unfortunately, there are other variants of DA, that cannot be so easily manipulated.

I have a separate map on the various DA options.

D6. Control of the simulation (if we are in it)

- Living an interesting life so our simulation will not be switched off
- Not letting them know that we know we live in a simulation
- Hacking the simulation and controlling it
- Negotiating with the simulators or praying for help

One of the risks is that we're inside a computer simulation created by a supercivilization with a purpose unknown to us, and that supercivilization can switch it off, or start testing inside it different variants of "doomsday". I'm working on another map dedicated to the simulation, where all these ideas will be discussed in more detail.

The share of simulations testing different doomsday scenarios can be quite large, as these simulations are necessary for any civilization spreading through the Universe and desiring to know what the number of other supercivilizations is. For this purpose it is necessary fir that supercivilization to carry out numerical simulation of the Fermi paradox, and, in particular, to find out how often civilizations are self-destructing.

However, if our civilization overcomes all risks within the simulation, it can still get shut down as it is no longer needed for the purposes of the experiment.

Bad plans

Bad plans are those plans that are actually better not to implement as they certainly increase the likelihood of a global

catastrophe. However, these ideas have been repeatedly expressed, and they may even be tried to be implemented, so it is important to list and criticize them.

Prevent x-risk research because it only increases risk

- Do not advertise the idea of man-made global catastrophe
- Don't try to control risks as it would only give rise to them
- As we can't measure the probability of a global catastrophe, it may be unreasonable to try to change the probability
- Do nothing

The essence of this proposal is to conceal the fact that new technologies bring new risks: we want to create every new tech sooner and get useful things from it. For example, to quicker obtain life extension through the development of biotech and nanotech but at the price of a small increase in risk of global catastrophe. Or gain a competitive advantage in the course of international confrontation or in finance.

But here we have the tragedy of the commons, because if many actors slightly raise a global risk for their personal gain, the total risk will grow much higher and make a catastrophe inevitable.

The following example is often presented: the research in the field of nanotechnology was largely frozen out of fear of "gray goo" after the Bill Joy's article.

It is also said that the idea of man-made disasters can inspire someone, and that person will become a super-terrorist. But this concealment does not work, as all the interested parties are already aware of the idea of global risks, basically from literature and movies. This idea is already well known. But the ways to prevent risks are much less known.

Anyway some ideas may be worth concealed or buried in the technical language or limited for exchange in a trusted experts network, as is in the case of the ideas to create bioweapons.

The premise that monitoring systems are creating new risks is true, but at the level of danger, when even a single bioterrorist can destroy all of mankind, some system of control is needed, or we are doomed.

If we do not deal with a possible disaster, it becomes inevitable.

And while we can't measure the exact probability of a global risk we could estimate the future survival time and also the frequency of smaller catastrophes.

Controlled regression

- Using a small catastrophe to prevent a large one (Willard

Wells)

- Luddism (Kaczynski): relinquishment of dangerous science
- Creating an ecological civilization without technology ("World made by hand", anarcho-primitivism)
- Limiting personal and collective intelligence to prevent dangerous science
- Radical antiglobalism and diversification into multipolar world (may raise probabilities of wars)

The idea of controlled regression, that is, lowering the level of technology, has repeatedly occurred in various forms. E.g., in one post-apocalyptic sci-fi story a world was described in which the death penalty was introduced to the inventors of the wheel.

If there are no hazardous technologies around, they will not create global risks. But if sustained regression has been achieved, humanity will soon die out by itself, like most previous earthly species. Or will again create tech because it cannot regress and have total a global monitoring system that would ensure its implementation in all parts of the Earth.

Moreover, the very achievement of regress requires certain dangerous acts.

For example, a nuclear war that would destroy the leading technology country in the world, can be an instrument of such a regression. But it would not only be a senseless crime, it may lead to the total extinction of mankind, or not achieve the stated objectives to stop progress for more than a few years. Or even to accelerate it in bad ways such as the creation of new types of weapons in the remaining countries. Theoretically, humanity can be in such a situation that small catastrophes will happen very often and it will prevent the creation of dangerous tech, but such a scenario is unlikely to be sustainable.

Another approach, which Kaczynski tried to implement, is targeted terrorism against individual scientists involved in the development of AI. He got life in prison. Large scale terrorism will entail drastic control measures that will balance it or result in arms race between terrorists and national security, which would produce even larger acts of violence and research in totalitarian control, and thereby accelerate existential risks. Luddism has no future. https://en.wikipedia.org/wiki/Neo-Luddism

Another idea is the creation of an environmental lifestyle in which mechanical work is replaced by manual work. This utopia is described in the novel "World made by hand. "https://en.wikipedia.org/wiki/World_Made_By_Hand

Another way of regression is the attempt to lower the intelligence of people so that they lose the ability to invent, with the help of a certain poison, a virus, or even through the destruction of

the system of universal education and tele-duping. But, of course, those methods will not work, or only reduce the total survivability of mankind.

Depopulation

- Natural causes: pandemics, war, hunger (Malthus)
- Extreme birth control
- Deliberate small catastrophe (bio-weapons)

Computerized totalitarian control

- Mind-shield – controlling dangerous ideation by means of brain implants
- Secret police that uses mind control to find potential terrorists and stop them

The idea that a population reduction may help to counter global risks is well known. Firstly, because a smaller population consumes fewer resources, and secondly, because it is easier to control a smaller population and fewer people in the world will be dangerous terrorists and gloomy supergeniuses.

The first idea was put forward by Malthus, who suggested that wars, famines and epidemics will naturally adjust the size of a too quickly growing population. However, a Malthusian catastrophe cannot result in human extinction.

Bill Gates offered to regulate the birth rate by reducing infant mortality and other soft methods. But the effect of such soft techniques can only be visible over periods of several decades, and during that time many of the exponential risks may occur.

The fact that billionaires expressed the idea to reduce the birth rate causes a reasonable fear that they have devised some more dangerous techniques to reduce human population. Such conspiracy theories could undermine any reasonable efforts in population regulation.

Declining birth rates may be organized technologically, through some form of biological or chemical weapons reducing fertility.

In general, the need for this will soon disappear because without it the world's total fertility rate has fallen sharply in recent years and is now only 2.35 births per woman, only slightly above the reproduction threshold, and continues to fall, thanks to education, city life and rising living standards.

In addition, regarding overpopulation as a problem prevents us from effectively seeking a cure for old age.

Choosing the way of extinction: UFAI

- Quick dying off is better
- Any super AI will have some memories about humanity
- It will use simulations of human civilization to study the probability of its own existence
- It may share some human values and distribute them throughout the Universe

Granted that the extinction is inevitable, we had better choose the way it will happen.

The worst case would be painful dying off because of a slow pandemic or radioactive contamination. (On the beach)

Immediate death resulting from vacuum phase transition or large scale asteroid collision seems to be a better option.

Immediate death by UFAI may be the best as it probably will keep intact the information about humanity, will run human simulations or preserve some of the human values or traits. But it could also be the worst if its goal system will include human torture (Roco Basilisk, and https://en.wikipedia.org/wiki/I_Have_No_Mouth,_and_I_Must_Scream)

Attracting good outcome by positive thinking

- Start partying now
- Preventing negative thoughts about the end of the world and about violence
- Assuming a maximum positive attitude "to attract" positive outcome

The plan was repeatedly expressed in various forms within the religious and magical community. It can take a form of collective meditations for the benefit of all. So far there is no scientific evidence that a private or collective intention to shape the future by a certain non-physical way will work.

The core idea of a feast in time of plague is to accept the inevitability of a global catastrophe, at the same time trying to realize as much as possible about personal values before it happens. In this case it is entertainment, but it may be other values. But as a result of no action a disaster can occur even earlier than expected.

Another idea in this line is that if everyone is engaged in personal entertainment, no-one will stage any dangerous scientific experiments, or attempt to take over the world. or arrange attacks. And everything will change for the better. But some people may find fighting for world domination a form of entertainment.

Conclusion

These plans of x-risks prevention may become a starting point for a productive discussion, which may result in some kind of an official law or an international roadmap to fight global risks.

This map is open for addition and I will constantly update it based on new ideas and considerations.

But as a based survey of exiting literature it is now the most complete, and the best ordered roadmap of the known methods of x-risks prevention.

Unfortunately, the situation in the world is deteriorating.

This map is part of a large project that will cover most futuristic topics: AI, life extension, other x-risks fields. The closest to this map is a map of typology of global catastrophic risks.

Other maps will be maps of AI failures levels, AI safety solutions, the casual structure of the global catastrophe, double scenarios of the global catastrophe.

Literature: